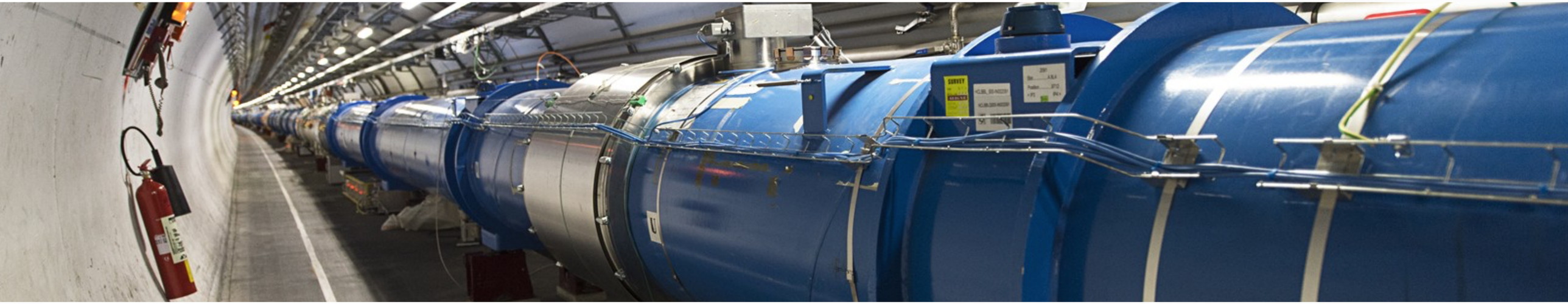
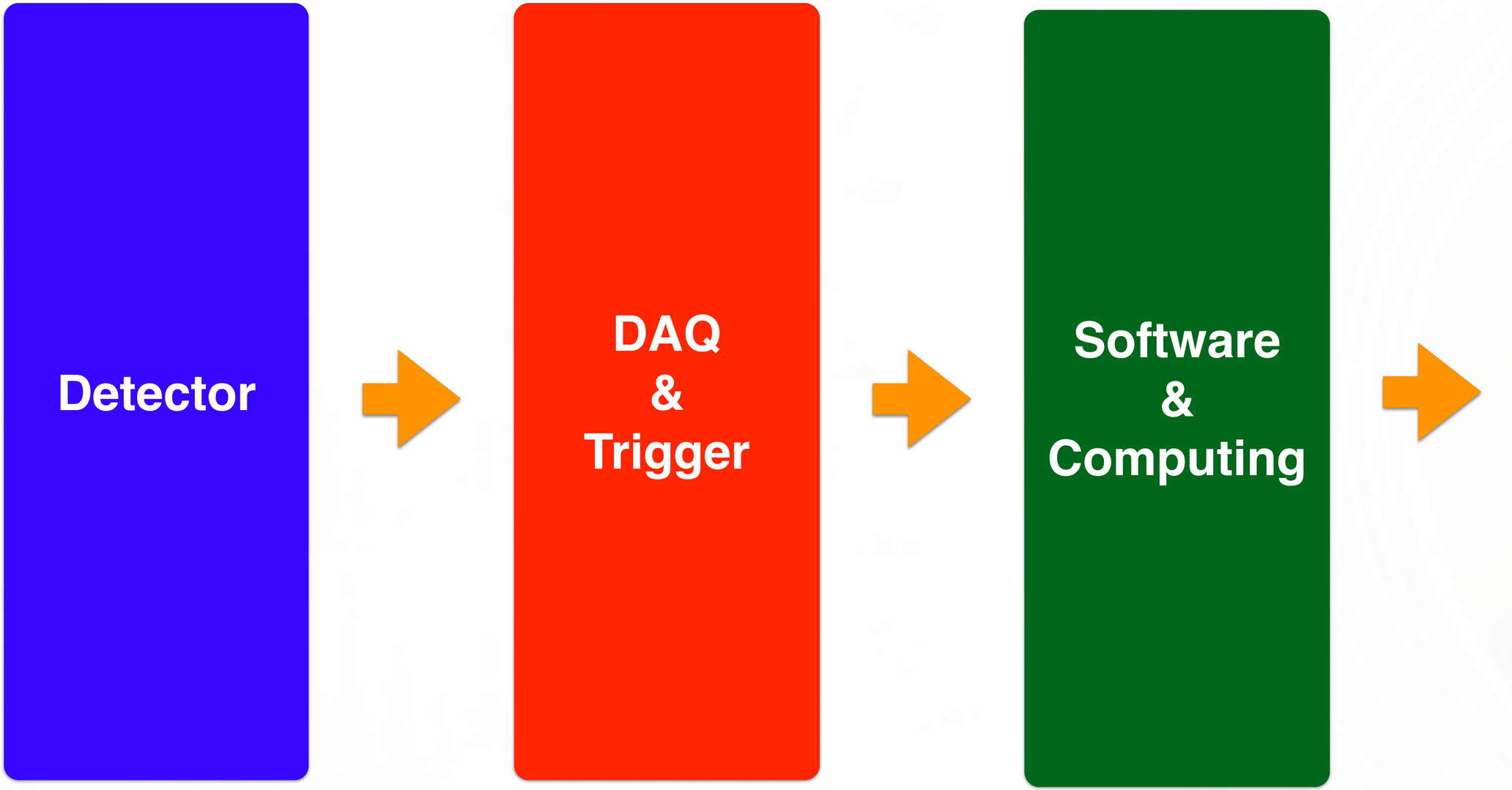


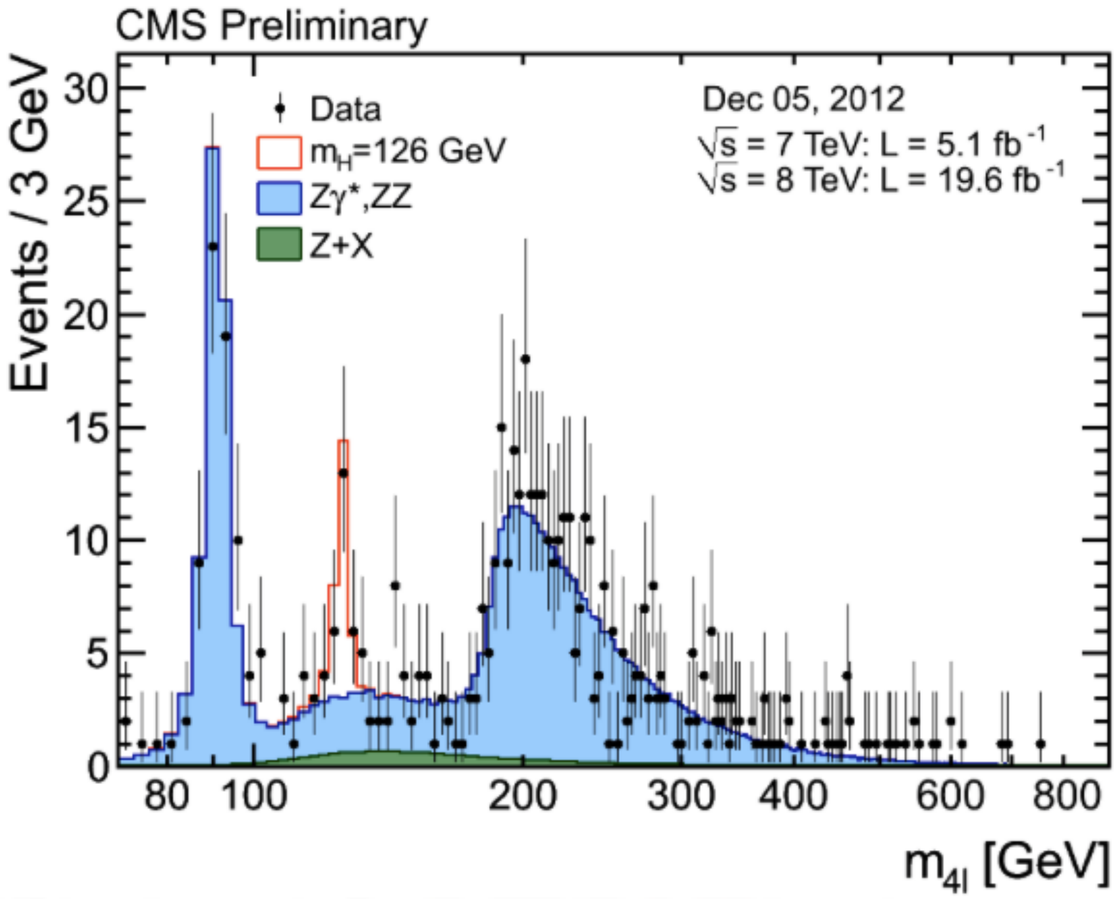
CMS Software and Computing in LHC Run 2 (and Beyond)

Matteo Cremonesi
FNAL

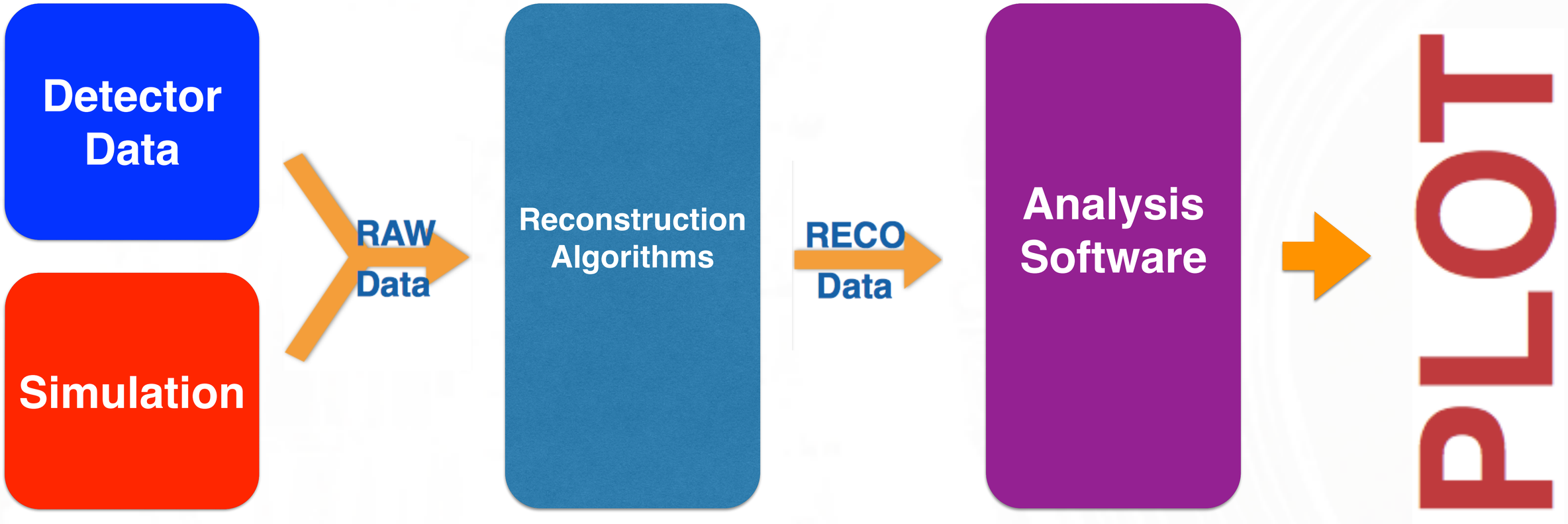
DPF - August 3, 2017

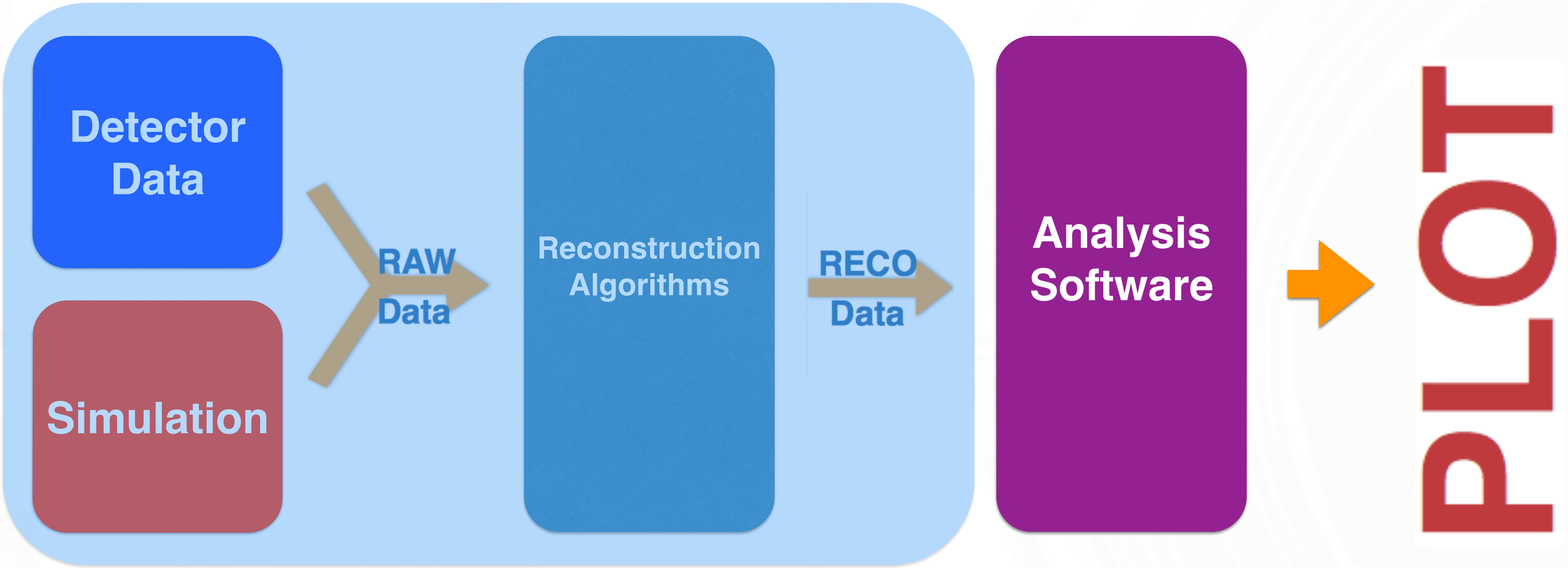

Detector
DAQ
&
Trigger
Software
&
Computing
SCIENCE

Experimental Particle Physics from Computing Perspective



- Detect particle interactions (data), compare with theory predictions (simulation)
  - **Black dots: recorded data**
  - **Blue shape: simulation**
  - **Red shape: simulation of new theory (in this case the Higgs)**


Detector Data
Simulation
RAW Data
Reconstruction Algorithms
RECO Data
Analysis Software
PLOT

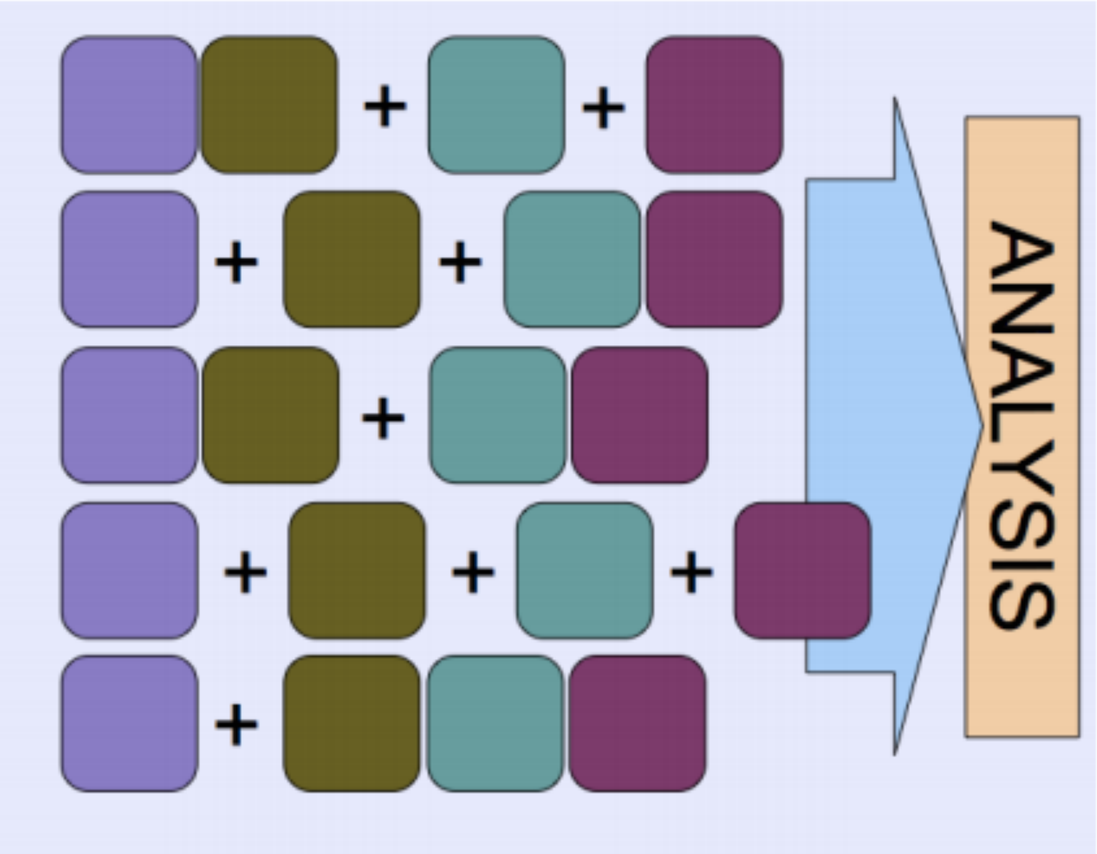


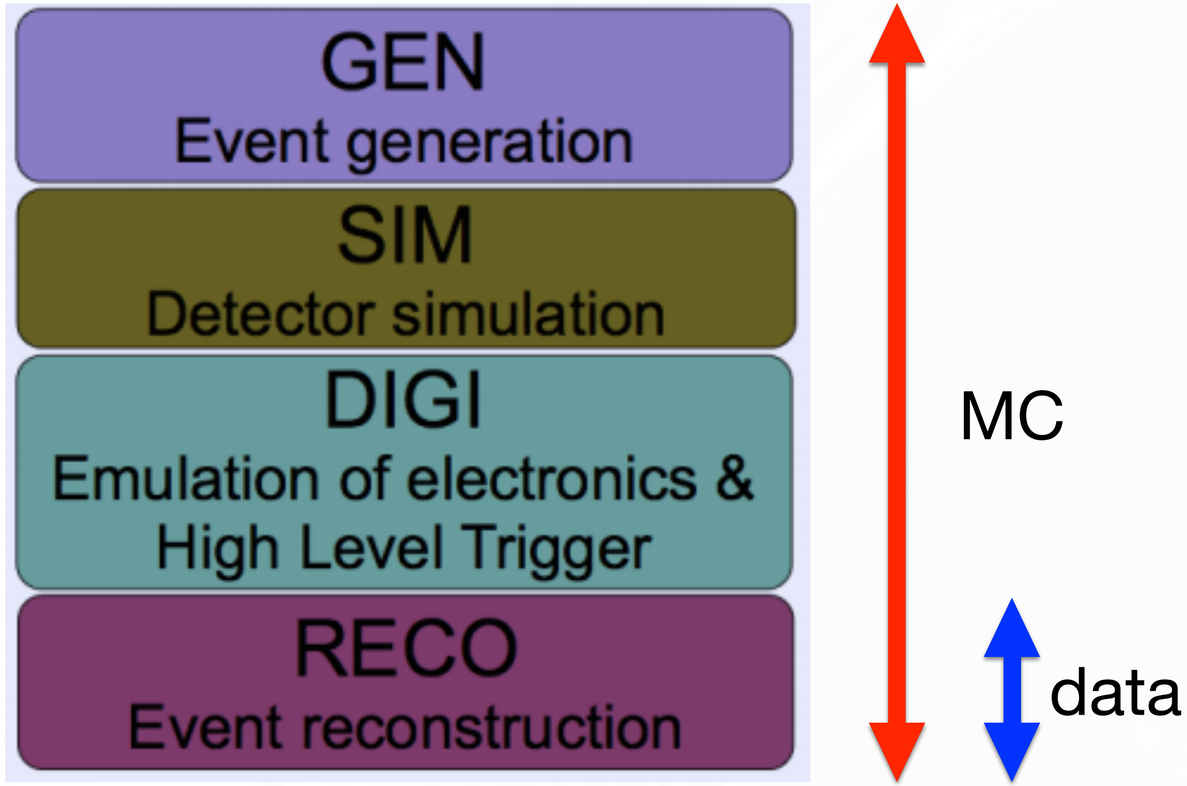
Central Production

Outline

- The challenge for central production
  - What is a workflow? How much work is there? Where can it run?
  - Description of the system needed to get all the work done
- Work assignment tool - more detailed description
  - Crucial for the efficient production of simulation and processing of detector data
  - Minimizes time to delivery of datasets for physics analysis => maximizes resource utilization

Request: Definition of Workflow





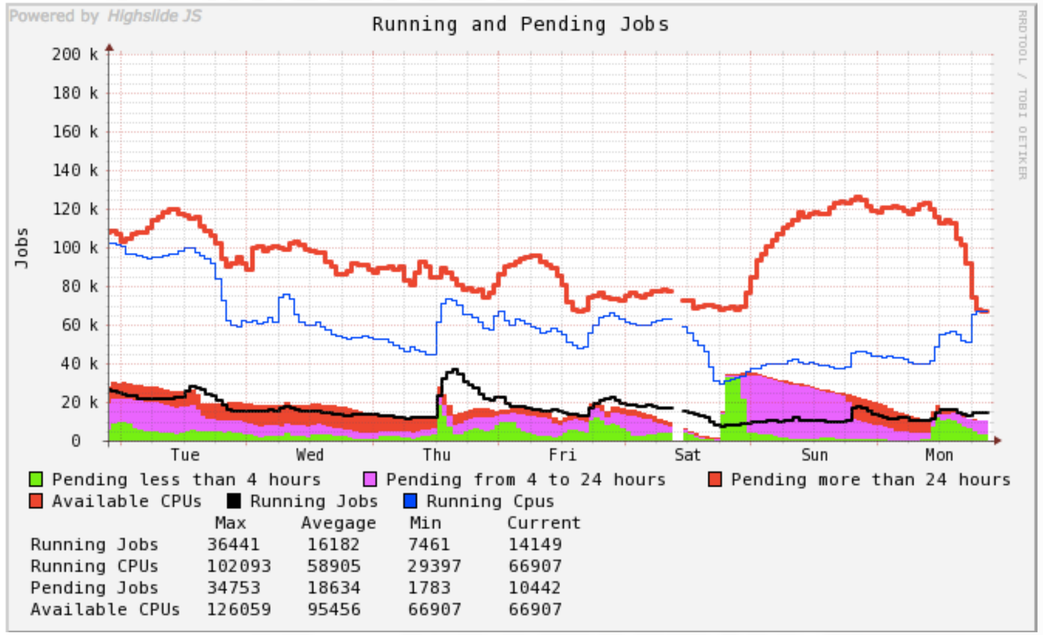
- abstract definition of processing and producing datasets
- converted into an actual sequence of jobs => production system
- defined by a set of algorithms, input, and output dataset

CMS Global Computing Grid

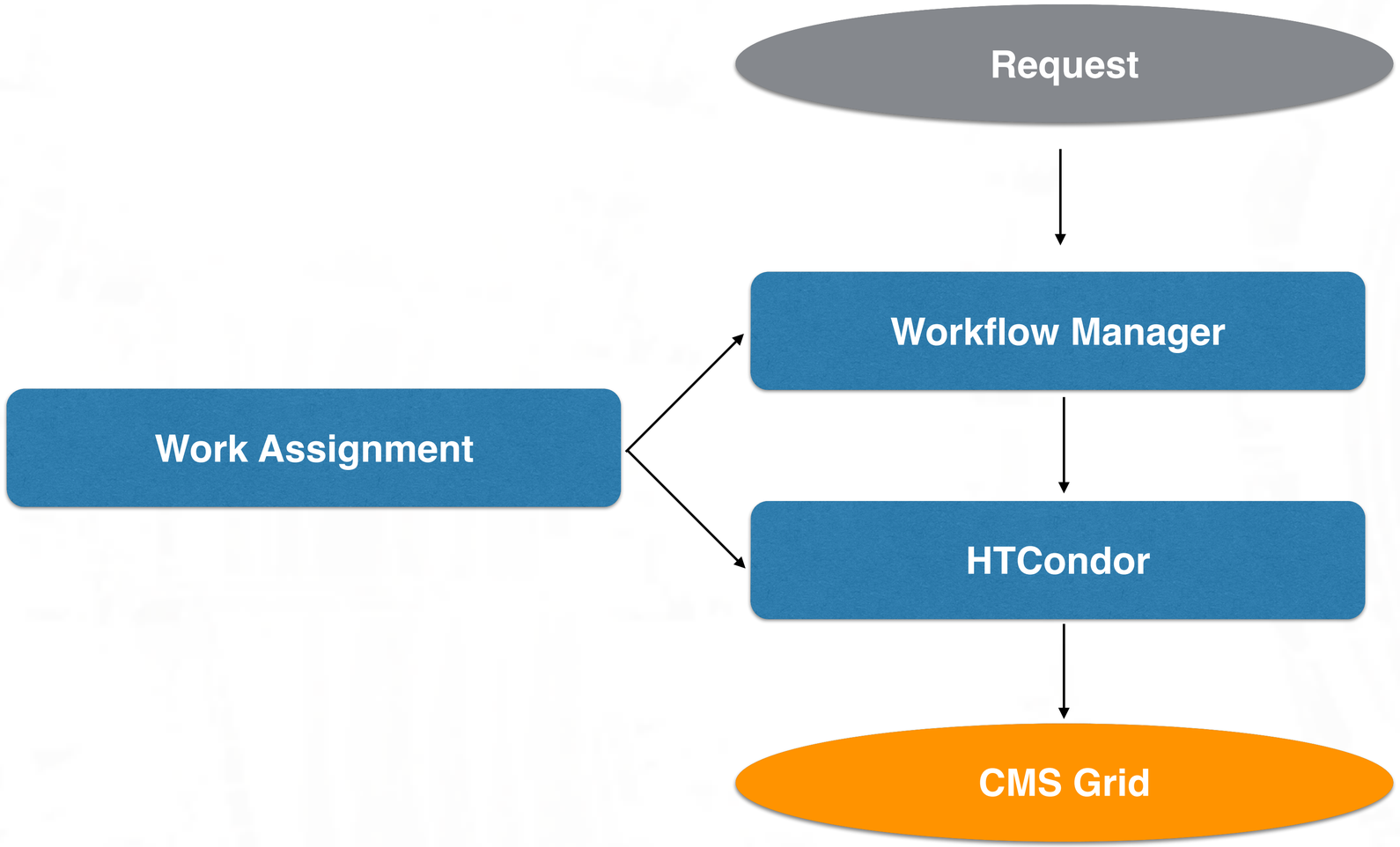
70+ sites, 150k+ CPU cores

Some Numbers

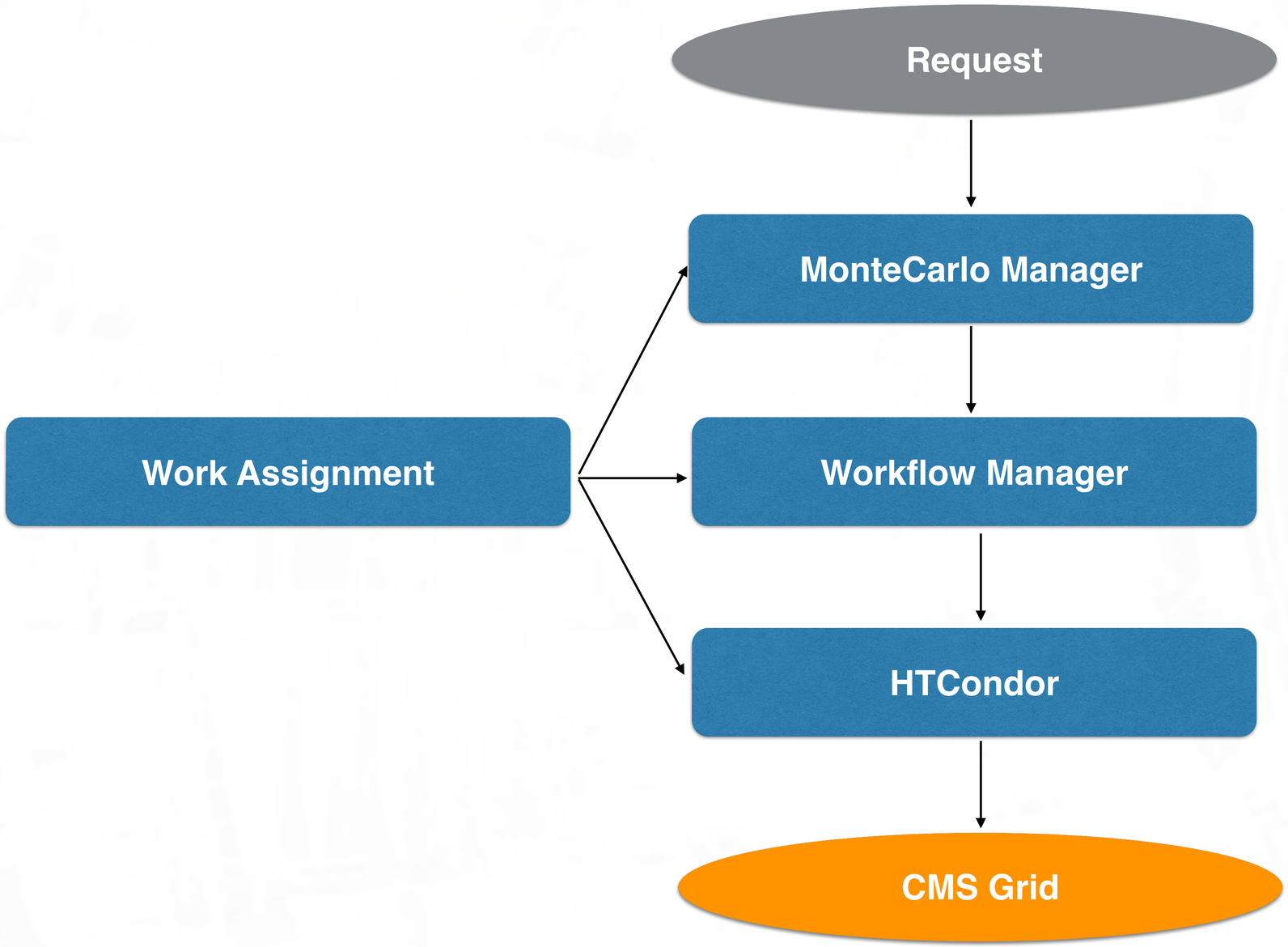
- Analyzing CMS data requires a large volume of simulation
  - Billions of events in 10s of thousands of datasets
- Requires a flexible and automated production system, needs to support at all times:
  - Up to 5k workflows in parallel
  - Up to 200k jobs pending,150k jobs running
    - Record: 200k concurrent jobs



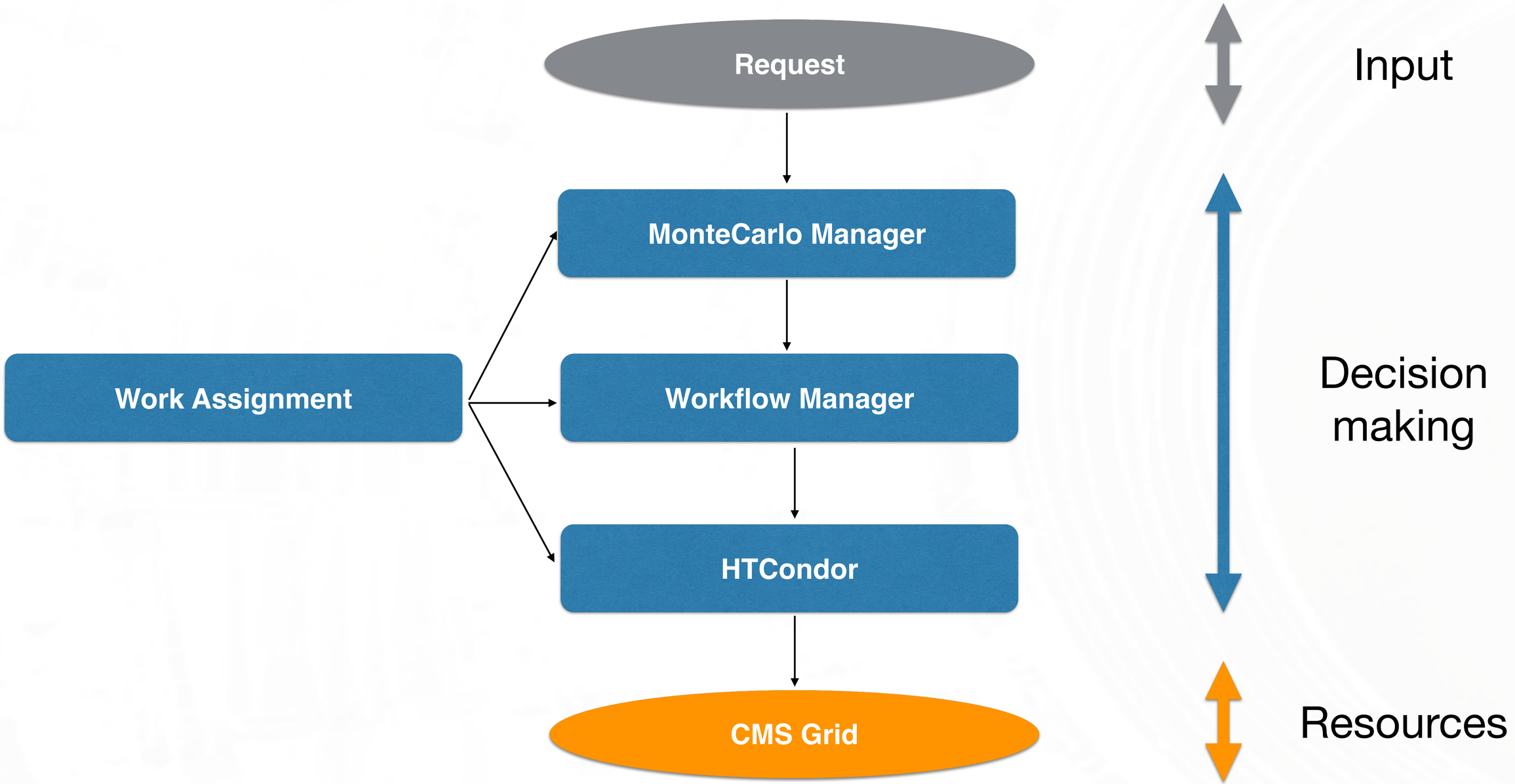
Data Processing at CMS



Simulation Processing at CMS



Simulation Processing at CMS



Some Technicalities of the System

McM (MonteCarlo Manager)

- Receive sample requests from the physics group.
- Inject consolidated workflows to production system.

ReqMgr (Workflow Manager)

- Receive assembled configuration from McM, prepare the full tree of processing towards the production of the final outputs.
- Split jobs according to workload specifications and data content and submit jobs to HTCondor.
- Resubmit certain types of failures.

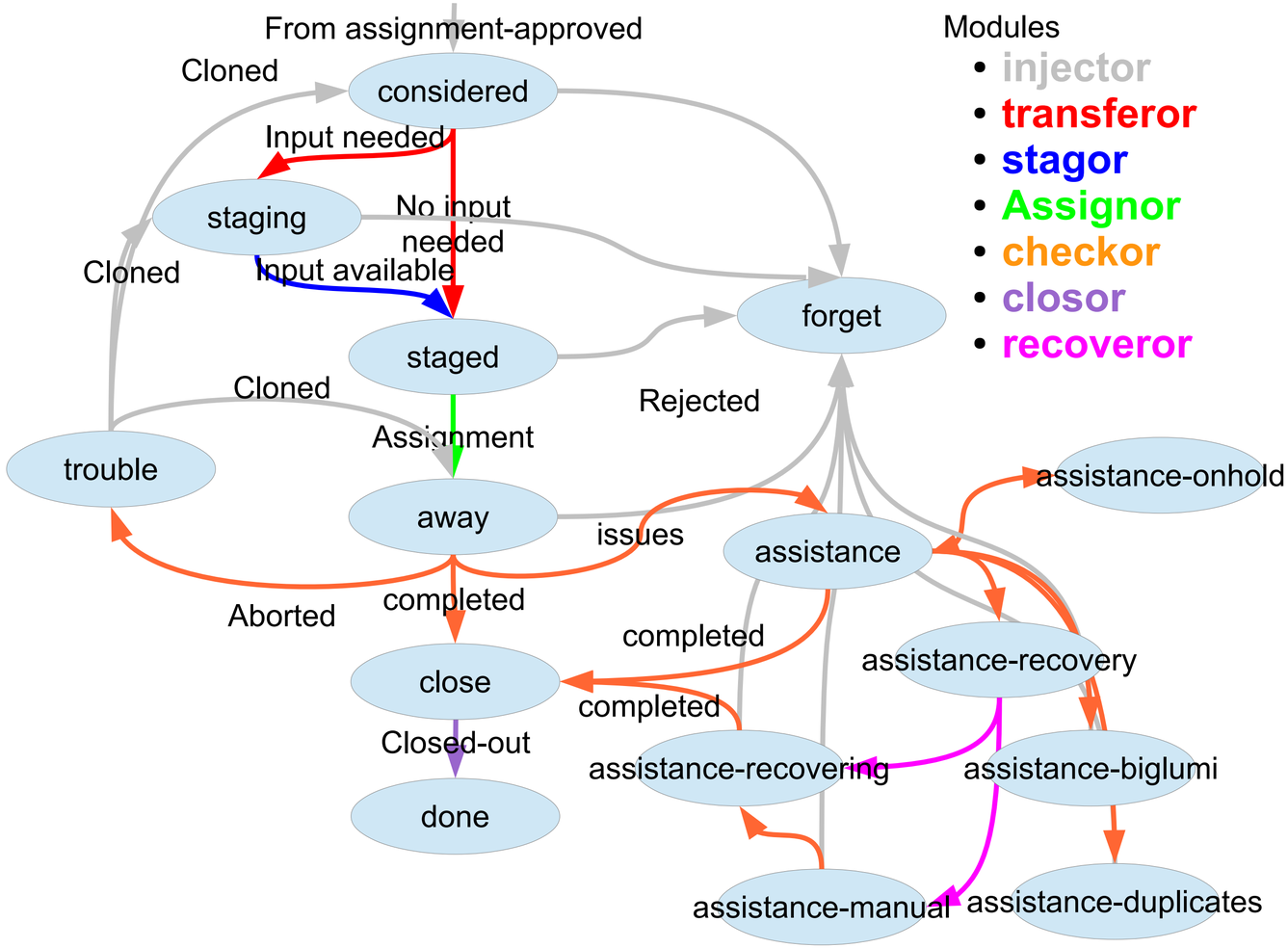
HTCondor

- Use shared resources between **analyzers** and **central production** in a global pool.
- Allow multi-core application, moving most workflows to 4+ threads

New!!!

Work Assignment: Unified

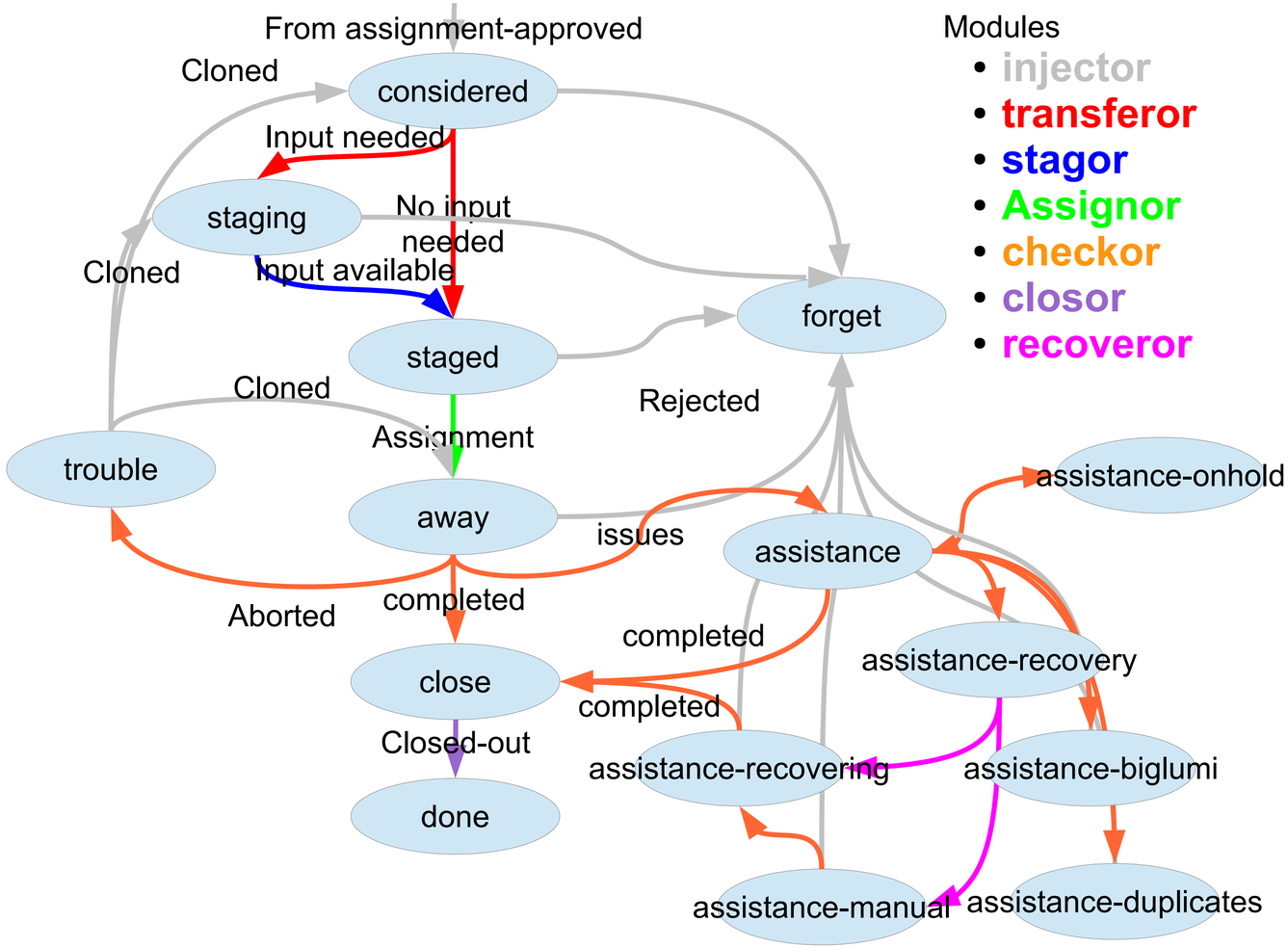
- A software to drive the workflows from the requester through ReqMgr and back to the requester.
- it solves a multi-dimensional matching problem: data location, available resources, etc.
- It does everything **automatically**
  - less effort needed
  - higher efficiency
  - optimized resource utilization



Work Assignment: Unified

Automation of transfer

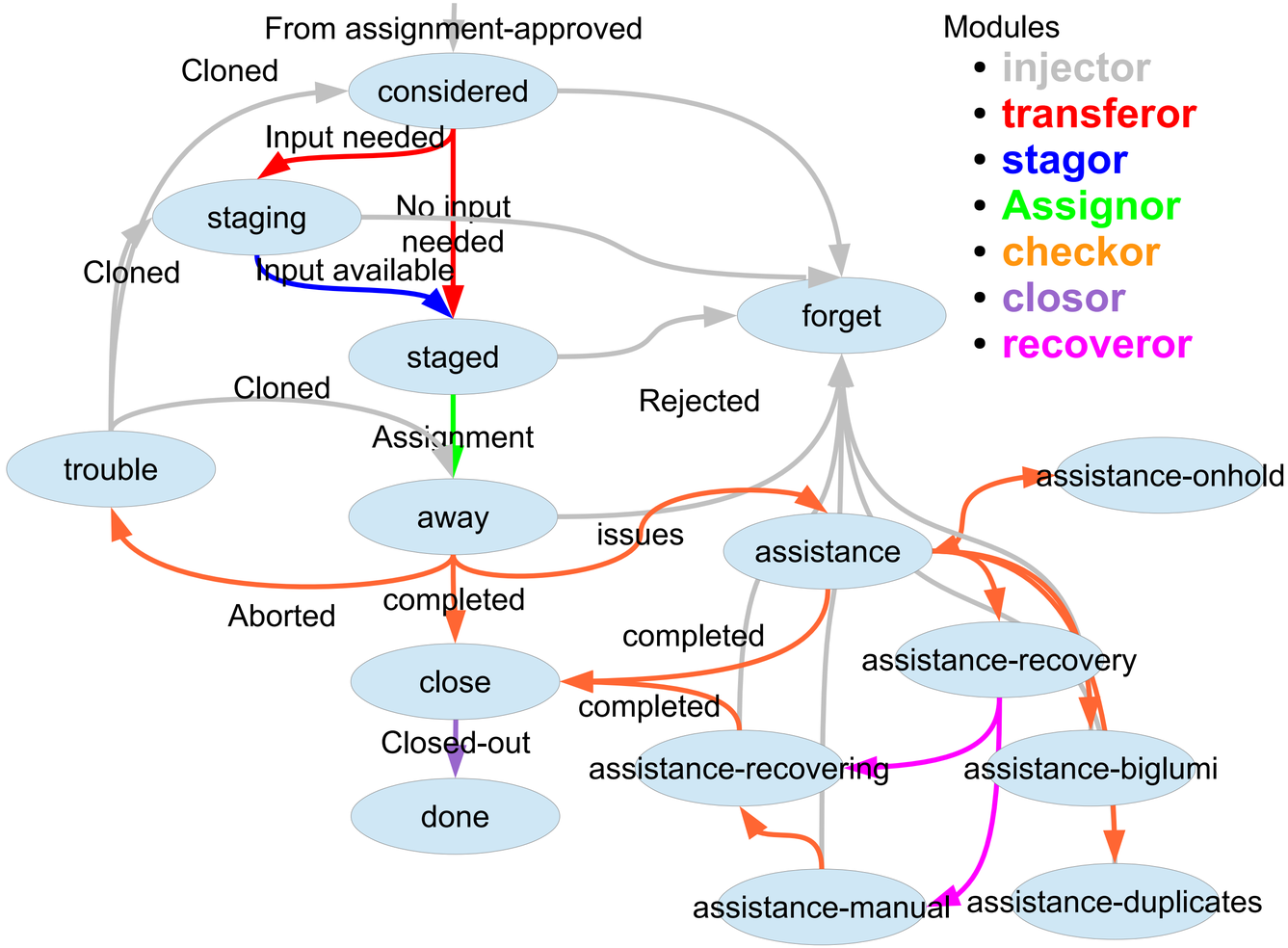
- parametrized number of copies of the input data to sites
- Destinations picked according to CPU pledge
- Monitoring of transfers



Work Assignment: Unified

Automatic assignment to as many sites as possible:

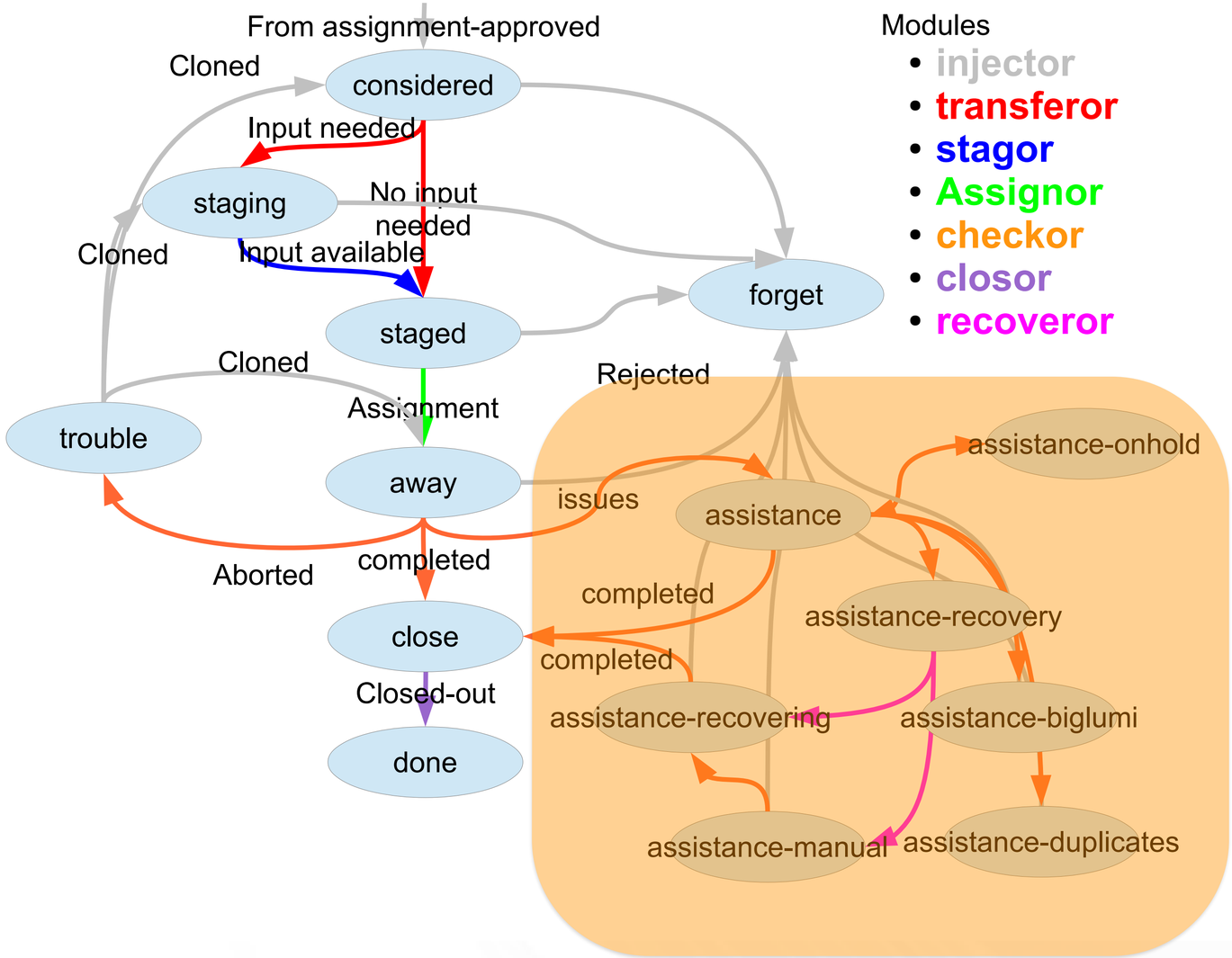
- Mostly homogeneous resource, but not all sites are equivalent (performance, policy, availability, size, ...)
- Thousands of workflows with heterogeneous requirements (CPU bound, I/O bound, high memory ,...)
- Balance job priority with site availability



Work Assignment: Unified

Automatic recovery

- Most workload are without issue (transfer, job failures, site issues, ...)
- Issues are dealt with increasing automation



Recent Developments on Unified

Overflow mechanism

- Site might come out of production status because of schedule intervention, emergency shutdown, intermittent failures
  - Workload backlog might develop on local site queue
- Mechanism to overflow to neighboring site
  - Quicken delivery with reliable remote read
- In future perspective, can be used to redirect jobs to resources becoming available

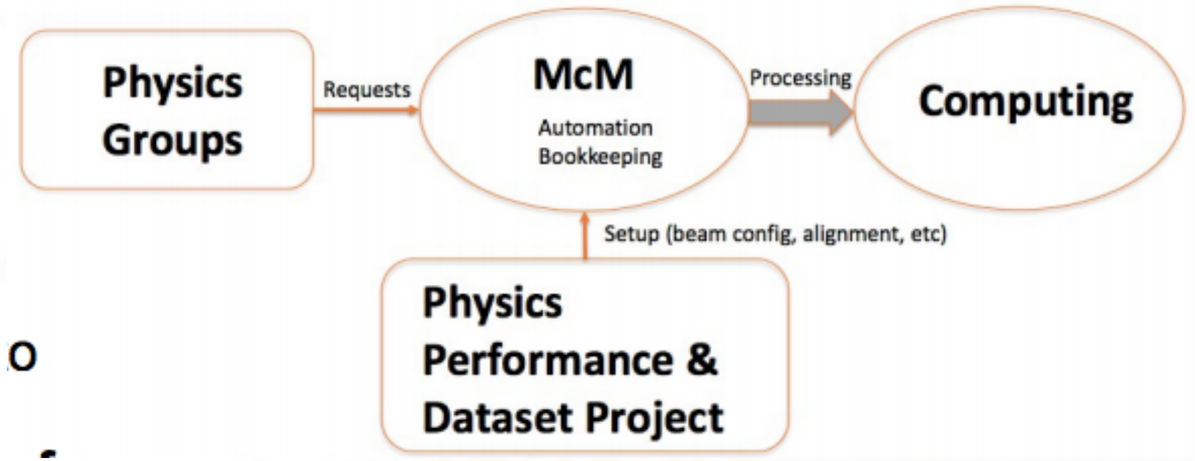
Conclusion

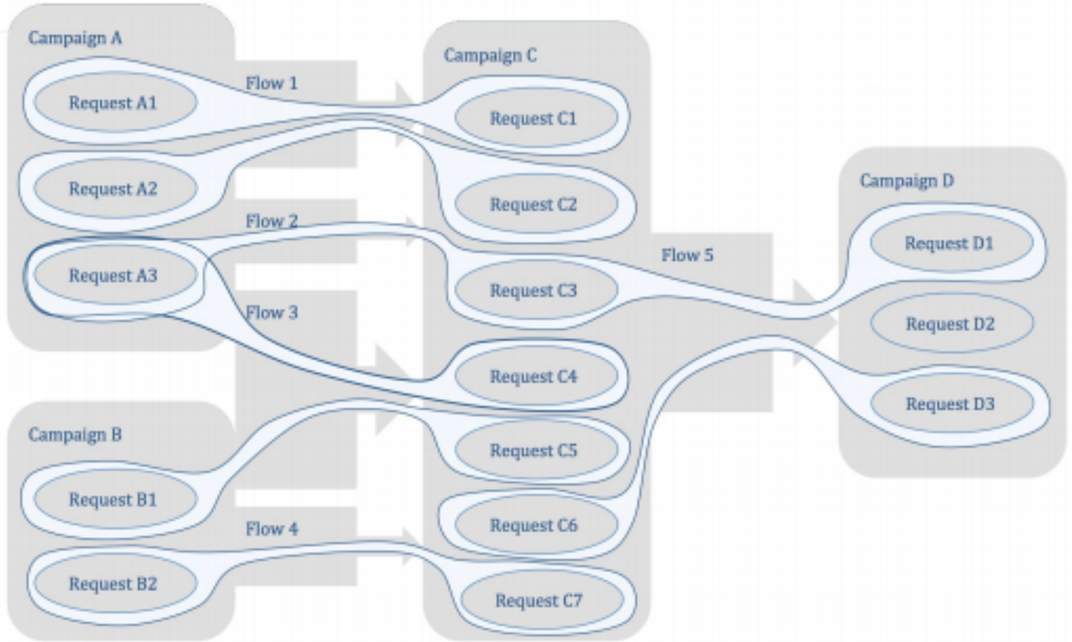
- CMS relies on a sophisticated infrastructure to process detector data and produce simulations
  - Without the timely and efficient delivery of thousands of samples CMS physics program would not be possible
- Workflow assignment tool instrumental in the success to deliver datasets to physics analysis in time
  - Supports large scale production and reprocessing for LHC Run II
  - Automates all steps of the production and processing cycle
  - Constantly working on improvements by learning from operation and investing in development

Backup

MonteCarlo Management: McM

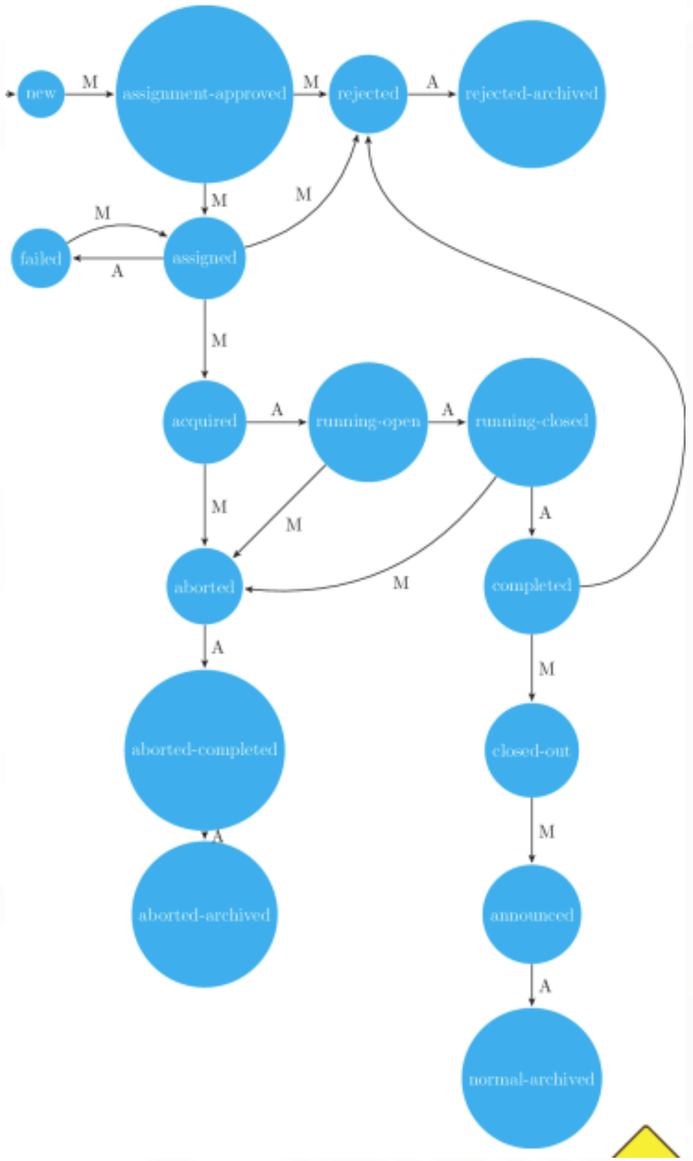
- Receive sample requests from generator contact person
- Inject consolidated workflows to production system
- CMS Software configuration and ingredients for production steps aggregated in campaigns
- Subsequent steps of production materialize in chains of campaigns
- Flow implement campaign modifiers
- Allow for complex chaining
- Flexibility for defining any specific request





Workflow Management: ReqMgr

- Receive assembled configuration from McM.
- Prepare the full tree of processing towards the production of the final outputs.
- Split jobs according to workload specifications and data content.
- Submit jobs to broker.
- Resubmit certain types of failures.
- Inject the produced data with parentage into book keeping system
- System composed by central request manager and multiple agents supporting high load
  - 5k workflows
  - 200k jobs pending
  - 150k jobs running



Job Broker: HTCondor

- Job broker that uses shared resources between analyzer and central production in a global pool.
- Use GlideIn mechanism:
  - Wrapper job: pilot running on site
  - Receive and execute trusted jobs
- Double stage of matchmaking
  - Jobs to resource (start pilots)
  - Jobs to pilots (claim pilots)
- Migrated for a large fraction to multi-core partitionable pilots
  - Allows multi-thread application, moving most workflows to 4+ threads
- Performances:
  - Record 200k concurrent jobs
  - Steady >150k job